Astronomical Data Analysis Software and Systems XVIII
ASP Conference Series, Vol. 411, © 2009
D. Bohlender, D. Durand and P. Dowler, eds.

A Cost-Benefit Study of Doing Astrophysics On The Cloud: Production of Image Mosaics

G. B. Berriman and J. C. Good
Infrared Processing and Analysis Center, California Institute of Technology, U.S.A.

E. Deelman and G. Singh
Information Sciences Institute, University of Southern California, U.S.A.

M. Livny
University of Wisconsin, Madison, U.S.A.

Abstract. Utility grids such as the Amazon EC2 and Amazon S3 clouds offer computational and storage resources that can be used on-demand for a fee by compute- and data-intensive applications. The cost of running an application on such a cloud depends on the compute, storage and communication resources it will provision and consume. Different execution plans of the same application may result in significantly different costs. We studied via simulation the cost performance trade-offs of different execution and resource provisioning plans by creating, under the Amazon cloud fee structure, mosaics with the Montage image mosaic engine, a widely used data- and compute-intensive application. Specifically, we studied the cost of building mosaics of 2MASS data that have sizes of 1, 2 and 4 square degrees, and a 2MASS all-sky mosaic. These are examples of mosaics commonly generated by astronomers. We also study these trade-offs in the context of the storage and communication fees of Amazon S3 when used for long-term application data archiving. Our results show that by provisioning the right amount of storage and compute resources cost can be significantly reduced with no significant impact on application performance.

1. Introduction

Clouds originated in the business world, and take advantage of modern web and networking technologies to offer businesses compute and storage facilities when they need them for as long as they need them (Amazon EC2). Briefly, clouds use virtualization technologies that enable applications to deploy a custom virtual environment suitable for a given application. Providers such as Amazon and Google charge applications for the use of their resources according to a fee structure. In this paper, we ask whether clouds are a powerful and cost-effective tool for astronomy applications. We have performed a cost-benefit analysis of using the Amazon Elastic Compute Cloud 2 (EC2)[1] to build astronomical im-

[1]http://aws.amazon.com/ec2/

age mosaics with the Montage image mosaic engine. A detailed description of this investigation is given in Deelman et al. (2008).

2. Production of Image Mosaics

Montage is a highly scalable and portable toolkit for assembling FITS images into science-grade mosaics that preserve the calibration and astrometric fidelity of the input images (Berriman et al. 2003). By design, the same code runs on desktops, clusters, grids, and supercomputers. It is available for download from `http://montage.ipac.caltech.edu`, and to date, there have been over 2,500 downloads. Montage has found wide applicability in astronomy in areas such as astronomical research, science-product generation, and education and public outreach.

Montage is written in ANSI-C for performance and portability, and deployed as a toolkit that performs the tasks needed to compute a mosaic:

- Find the input images that lie within the footprint of the output mosaic on the sky
- Reproject the input images to the required projection of the output mosaic
- Rectify the background radiation, which varies across the input images, to a common level across all images
- Co-add the reprojected and rectified images to form the output mosaic.

Each of these tasks generally take a few minutes to run, and the output from one module becomes the input to the next module. Montage is a data-intensive application and it needs to run on a resource-rich environment where storage resources are assured.

2.1. Design of Cost-Benefit Study

The study used the Gridsim tool[2] to simulate processing on the Amazon EC2 cloud of three mosaics of M17 of sizes 1, 2 and 4 square degrees, comprising respectively 203, 731 and 3027 tasks, under the Amazon EC2 fee structure that was current in the spring of 2008:

- \$0.15 per GB-Month for storage
- \$0.1 per GB for transferring data in
- \$0.16 per GB for transferring data out
- \$0.1 per CPU-hour for computing.

Amazon EC2 levies no charge for accessing data stored on its storage systems by tasks running on its compute resources. The processing was simulated on 32-bit 1.2 GHz Xeon processors with 1.7 GB memory running under Red Hat Enterprise Linux. The study excluded the costs of setting up the virtual image needed to process these mosaics. The virtual image in this example included the Montage application and the Pegasus workflow tools[3], needed to manage and schedule the processing.

---

[2]http://www.gridbus.org/gridsim/

[3]http://pegasus.isi.edu/

The sections below describe the costs, normalized to the cost per second, of producing mosaics for three use cases: using the cloud for occasional needs, with input data stored outside the cloud; using the cloud to provide all computing needs, with the input data stored outside the cloud; and using the cloud to provide all computing needs, with the data stored inside the cloud.

3. Results of the Cost-Benefit Study

3.1. Use Case 1: Occasional Requests with Data Stored Outside the Cloud

For all three mosaic sizes, CPU costs predominate over all other costs. For example, for the 1 square degree mosaics, the CPU cost rises from \$0.8 for 1 provisioned processor to \$8 for 128 processors. That is, the cost of provisioning more processors overcomes the decrease in computing time: the only compelling reason for provisioning more processors is to return the mosaic quickly. The transfer costs into the cloud remain constant at \$0.08, while the storage costs decline from \$0.1 for one processor, to \$0.005 for 128 processors. The explanation of these costs is straightforward. While the same amount of data must be transferred into the cloud regardless of the number of processors provisioned, the data are stored for less time as the execution time declines with the number of processors, approximately by a factor 12 as the number of processors increases from 1 to 128.

3.2. Use Case 2: Use Clouds for All Computing Needs, with Data Stored Outside the Cloud

In this case, processors are requisitioned and used only when jobs need them. The data-management costs assume greater importance than in the first use case, depending on how the data are managed. Consider data management costs for the 1 square degree mosaic for three methodologies:

- Stage all the input data for a task to the compute resources, execute the tasks, then download the output and delete all the data. This model applies to the case where there is no shared storage (remote I/O mode).
- When the compute resources have access to shared storage, it can store the intermediate files produced by the tasks, which in turn are the input to subsequent tasks running on another resource. Only when all tasks are completed are the intermediate files complete (regular mode).
- Employ dynamic cleanup, where intermediate files are deleted as soon as the workflow no longer needs them (dynamic cleanup mode).

Data management costs are highest for the remote I/O mode, and are comparable to the processing costs. For a 1 square degree mosaic with 128 processors provisioned, the processing and data transfer costs are roughly \$0.5. For the 4 square degree case, the corresponding costs are \$7. The regular and dynamic cleanup modes entail much cheaper data management costs than remote I/O For all mosaic sizes, their data management costs are a factor of ten less than the computational costs. This is despite the fact that the efficiency of dynamic cleanup offers the lowest data transfer costs, and reduces storage costs to half those of the regular mode. The data management costs in both modes are much

lower than computation costs, and so dynamic cleanup does not offer substantial cost benefits to the user.

3.3. Use Case 3: Use Clouds for All Computing Needs, with Data Stored Outside the Cloud

The cost of storing data on the cloud are sufficiently high that they only offer benefits to users if usage is intensive. For example, the cost of storing all three bands of the 2MASS All Sky Image Atlas, a total of 12 TB, is $1,800 per month on the Cloud. Processing a 1 degree mosaic 2MASS and delivering it to the user costs $2.22 with the input data outside the cloud, and $2.12 with the input data inside the cloud. To overcome the storage costs, users would need to request at least 18,000 mosaics per month.

4. Conclusions

Under the current Amazon EC2 cost structure:

- Processing costs more than data transfer and data management for the data-intensive mosaic application.
- The cloud is cost-effective for generating mosaics if data are transferred to the cloud.
- The cost of storing large image data sets on the cloud remains prohibitive unless use is intensive.

Acknowledgments. This work was funded by the National Science Foundation under Cooperative Agreement OCI-0438712 and grant number CCF-0725332. Montage was funded by the NASA Earth Sciences Technology Office Computing Technologies (ESTO-CT) program, under Cooperative Agreement Notice NCC 5-6261. It is now maintained by the NASA IPAC Infrared Science Archive, funded by NASA under contract to the Jet Propulsion Laboratory.

References

Berriman, G. B., et al. 2003, in in ASP Conf. Ser. 314, ADASS XIII, ed. F. Ochsenbein, M. Allen, & D. Egret (San Francisco: ASP), 593

Deelman, E., Singh, G., Livny, M., Berriman, B., Good, J. 2008, "The Cost of Doing Science on the Cloud: The Montage Example," Proceedings of Super Computing 2008, Austin, Texas